IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2019, accepted August 11, 2019, date of publication August 21, 2019, date of current version September 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936709

Improving Human Pose Estimation With Self-Attention Generative Adversarial Networks

XIANGYANG WANG, ZHONGZHENG CAO, RUI WANG, ZHI LIU, AND XIAOQIANG ZHU
School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

Corresponding author: Rui Wang (rwang@shu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771299.

ABSTRACT Human pose estimation in images is challenging and important for many computer vision applications. Large improvements in human pose estimation have been achieved with the development of convolutional neural networks. Even though, when encountered some difficult cases even the state-of-the-art models may fail to predict all the body joints correctly. Some recent works try to refine the pose estimator. GAN (Generative Adversarial Networks) has been proved to be efficient to improve human pose estimation. However, GAN can only learn local body joints structural constrains. In this paper, we propose to apply Self-Attention GAN to further improve the performance of human pose estimation. With attention mechanism in the framework of GAN, we can learn long-range body joints dependencies, therefore enforce the entire body joints structural constrains to make all the body joints to be consistent. Our method outperforms other state-of-the-art methods on two standard benchmark datasets MPII and LSP for human pose estimation. Our code is available at: https://github.com/idotc/Hg-SAGAN.

INDEX TERMS Human pose estimation, convolutional neural networks, stacked hourglass networks, self-attention GAN.

I. INTRODUCTION

Human pose estimation (HPE) aims to predict the locations of body joints from input images. It is fundamental for some other computer vision applications such as action recognition [1]–[3], human-computer interaction and video surveillance. The most recent methods for human pose estimation take advantage of convolutional neural networks (CNNs) to drastically improve the performance on standard benchmarks [8]–[10], [12]–[14].

Despite of these great progresses, there still exist some challenging cases, such as ambiguities caused by occluded body joints, invisible joints, nearby persons and clutter backgrounds, where even the state-of-the-art models may fail to predict the body joints correctly. The main reasons lie in: 1) these ''hard'' joints cannot be simply recognized based on their appearance features only; 2) these ''hard'' joints are not explicitly addressed during the training process [12].

One of the straightforward and efficient ways to handle these ''hard'' cases maybe combining body joints structural constraints into the training process to make the predicted pose plausible. GAN (Generative Adversarial Networks) [15] has been applied to learn the structural constrains of human body parts by adversarial training [10], [11].

However, there are problems with existing GAN based pose estimation models. Since traditional convolutional GANs can only learn the spatially local constraints, previous GAN based HPE methods [10], [11] still cannot fully tackle these challenging cases when more complex body joints occlusion and crowded backgrounds occur.

Recently, Zhang and Goodfellow et al propose the Self-Attention Generative Adversarial Networks (SAGANs) [16], which introduce a self-attention mechanism into convolutional GANs. The self-attention module is complementary to convolutions and is capable of modeling long-range, multi-level dependencies across image regions. With self-attention, the discriminator can more accurately enforce complicated geometric constrains on the global image structure [16], which leads the generator to produce holistic consistent images.

Motivated by SAGANs, in this paper, we propose to apply self-Attention GAN to further improve the performance of human pose estimation. With attention mechanism in the framework of GAN, we can learn long-range

The associate editor coordinating the review of this article and approving it for publication was Varuna De Silva.

body joints dependencies, therefore enforce the entire body joints structural constrains making all the body joints to be consistent. We evaluate the proposed approach on two HPE benchmarks, MPII and LSP. Experimental results show that our approach outperforms state-of-the-art methods.

II. RELATED WORK

A. HUMAN POSE ESTIMATION

Human pose estimation (HPE) is a challenging problem due to the large variations in configuration and appearance of body parts. Early works often tackle such problems by graphical models with handcrafted image features [17]–[19].

Similar as many other vision tasks, the progress on human pose estimation has been greatly advanced by deep learning [20], [21], since Convolutional Neural Networks (CNNs) have the powerful ability to learn rich convolutional feature representations. Before CNNs were applied for HPE, the performance of previous works on the MPII benchmark [22] was only about 40% PCKh@0.5 [17]. CNNs pioneer works surprisingly improve it to about 80% [4], [5]. During the later three years, till now, it has achieved to more than 90% [8]–[10], [14]. The mAP (mean Average Precision) metric on more recent and challenging COCO human pose benchmark [23] has been increased from 60.5 (COCO 2016 Challenge winner [6], [7]) to 72.1 and recently 78.1 (COCO 2017 and 2018 Challenge winner [12]) [13].

1) SINGLE PERSON POSE ESTIMATION

DeepPose [4] is the first deep learning based approach for human pose estimation, which takes pose estimation as a body keypoints regression problem using Convolutional Neural Networks. Latter methods mostly predict heatmaps that characterize the probabilities of each keypoint at different locations [5]. The exact location of a keypoint is further estimated by finding the maximum in an aggregation of heatmaps. Heatmap-based methods better leverage the distributed properties of convolutional networks and are more suitable for training human pose estimation models [10].

Some works combine graphical models with CNN. Tompson *et al.* [5] apply MRF (Markov random field) as a post-processing step, while others embed deformable mixture of parts [24] or CRF (Conditional random field) [25] into the network for end-to-end learning. Convolutional Pose Machines (CPM) [6] and Stacked Hourglass Network (Hourglass) [8] achieve state-of-the-art performance without hand designed priors or graphical model-style inference. Both CPM and Hourglass employ a multi-stage scheme, using intermediate supervision to produce increasingly refined heatmaps for joints locations through different stages. The design of Stacked Hourglass Network [8] is motivated by FCN (Fully Convolutional Networks) [26] and ResNet [27]. Its powerful and well-designed architecture consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference [26]. Features are processed across all scales and consolidated to best capture the various spatial relationships associated with the body. Each Hourglass module contains several residual modules [27].

After that, most of the recent single person pose estimation methods are based on Hourglass, and try to improve Hourglass either by multi-scale feature pyramids (PyraNet) [9] or by multi-stage refinement [10], [14].

Feedback Networks [28] improve Hourglass by adding feedback connections. Inspired by DenseNet [29], Tang *et al.* [30] add dense skip connections into the residual blocks across different Hourglass modules. Yang *et al.* propose the PyraNet [9] with a Pyramid Residual Module (PRMs) to enhance multi-scale invariance in CNNs. In fact, PyraNet is an improvement of Hourglass by extending the residual block to multi-scale of pyramids. If the level of pyramids is 1, it is exactly the original Hourglass networks. Ke *et al.* [31] also intend to improve Hourglass by multi-scale CNN features. Additionally, they incorporate a structural loss into the training process.

In [10], [11], Generative Adversarial Networks (GAN) are utilized to human pose estimation. Chou *et al.* [10] put the Hourglass networks into the GAN framework with self-adversarial training, where the generator and discriminator are all hourglass modules (We denote their work as Self-GAN). The generator predicts human pose, and the discriminator acts as a judger to enforce structural constraints of human body joints.

Tang *et al.* [14] propose a Deeply Learned Compositional Model (DLCM) for human pose estimation. It exploits CNNs to learn the compositionality of human bodies. The network has a hierarchical compositional architecture and bottom-up/top-down inference stages. In the bottom-up stage of DLCM, Hourglass is used to predict human pose. And subsequently the top-down stage plays the role to refine the predicted pose in bottom-up stage.

DLCM and Self-GAN are very similar. The generator in Self-GAN is similar to the bottom-up stage in DLCM to predict human pose, the discriminator in Self-GAN and the top-down stage in DLCM both act as pose refinement. We can see that all the above mentioned recently developed works are indeed improved hourglass.

2) MULTI-PERSON POSE ESTIMATION

The more practical problem is multi-person pose estimation, which is to estimate poses of multiple people in one image. There are two types of methods for multi-person pose estimation, bottom-up and top-down. Bottom-up methods first locate keypoints for all persons in the image and then group joints candidates for each person. Such as, DeepCut, DeeperCut [32], [33] and Openpose [7]. DeepCut and DeeperCut [32], [33] use CNNs (VGG [35] or ResNet [27]) to generate keypoint candidates and then run integer linear programming (ILP) to group them for each person. Cao *et al.* propose the Openpose [7], which is based on CPM [6] to simultaneously learn multi-person joints locations and their associations via Part Affinity Fields (PAFs), and then uses a

greedy algorithm to group the joints that belong to the same person.

Top-down methods [12], [13], [34], [36] are also two-stage methods, they first detect each person in the image, and then estimate the pose for each single person. Fang *et al.* propose the RMPE (AlphaPose) [36], which uses SSD [37] (or Faster RCNN [38]) to detect persons in the image and then utilizes Hourglass [8] (or PyraNet [9]) to estimate poses for each person. Chen *et al.* propose Cascaded Pyramid Network (CPN) for Multi-Person Pose Estimation [12]. CPN first applies Mask RCNN [34] to detect persons, and then designs the CPN for pose estimation of each person. CPN involves two subnetworks: GlobalNet and RefineNet. GlobalNet is based on ResNet backbone. Based on the feature pyramid representation generated by GlobalNet, RefineNet serves to refine the ''hard'' keypoints.

3) POSE TRACKING

The more challenging task is simultaneous pose estimation and tracking [13], [39] or pose estimation in videos [40], [41]. Luo *et al.* adopt Long Short-Term Memory (LSTM) to impose geometric consistency among video frames while using CPM [6] to estimate person pose in images. 3D human pose estimation is also very important for practical applications, such as virtual reality or augmented reality [42], [43]. But 3D human pose estimation is based on 2D. Usually, 2D pose must be estimated first and then extended to 3D.

In summation: (1) Recent state-of-the-art human pose estimation methods are either improved hourglass networks [9], [10], [14], [84], or take ResNet as their backbone [12], [13], [44]; (2) Among these tasks, 2D single person pose estimation is the basis. In this paper, we focus on 2D single person pose estimation from RGB images.

B. GENERATIVE ADVERSARIAL NETWORKS

Recently, Generative Adversarial Networks (GAN) has been applied to various computer vision tasks, such as image super-resolution [59], image inpainting [60], object detection [61], person image synthesis [62], person Re-identification [63], and human pose estimation [10], [11]. GAN is first proposed by Goodfellow *et al.* [15], which can generate natural images such as human faces and indoor scenes. It consists of generator and discriminator. The generator generates images to fool the discriminator, while the discriminator tries to distinguish the fake one from the real. In this way, the adversarial training can help generator to improve its product increasingly. The training of GANs may be unstable and sensitive to the choice of hyper-parameters.

Researches on GAN may be considered mainly from three perspectives: (1) some works try to improve the training of GAN; (2) some works shine light on GAN theoretic analysis; (3) most of the works exploit various GAN applications.

Radford *et al.* [45] introduce DCGAN with all convolutional architecture. They eliminate the fully connected layer and employ batch normalization to prevent from losing diversity, i.e., model collapsing. Arjovsky *et al.* [46] propose Wasserstein GAN (WGAN), which uses Wasserstein distance to replace the original loss function in GAN and solves the unreliable gradient problem. WGAN satisfies the K-Lipschitz constraint by weight clipping, which pushes weights towards two extremes of the clipping range and is hard to tune the clipping parameters. Gulrajani *et al.* [47] improve the training of Wasserstein GANs by replacing the weight clipping strategy with gradient penalty. Gradient penalty is an additional term in the loss function that directly enforces the discriminator's gradient norm around K. EBGAN [48] uses autoencoders as discriminators. It aims to match the autoencoder loss distribution instead of matching the data distribution. Based on EBGAN and proportional control theory, BEGAN [49] introduces an equilibrium term to balance the discriminator and the generator. It also provides a convergence measure to determine if the model has collapsed or reached its final state.

Recently, Miyato *et al.* propose the Spectral normalization GAN (SNGAN) [50] by limiting the spectral norm of the weight matrices in the discriminator in order to constrain the Lipschitz constant of the discriminator function. Zhang *et al.* [16] apply spectral normalization to the GAN generator to improve training dynamics.

Gu et al. utilize Optimal Transportation and Monge-Ampere equation to theoretically interpret deep learning and GAN [51]– [53]. They show the intrinsic relations between optimal transportation and convex geometry, the generator calculates the transportation map while the discriminator computes the Kantorovich potential [52].

Initially, GAN is used to generate synthetic images from input noises [15]. With rapid development in recent years, GAN has been able to generate amazing perfect images. Zhu *et al.* propose the state-of-the-art CycleGAN [54] to learn to translate an image from a source domain to a target domain in the absence of paired examples. CycleGAN learns bidirectional mappings between source and target domain with adversarial and cycle consistency losses to enforce the translation to be consistent. Bansal *et al.* propose the Recycle-GAN [55] for unsupervised video retargeting that enables the transfer of sequential content from one domain to another while preserving the style of the target domain. StarGAN [56] can perform image-to-image translations for multiple domains using only a single model. GANimation [57] enables continuous facial animation. Vid2vid [58] implements Video-to-Video Synthesis with GAN.

C. ATTENTION MODELS

Attention mechanism [64], [65] allows the model to differentiate irrelevant information so as to focus on the most relevant part of images or features as needed. Attention mechanism has been proven effective and successfully applied in many computer vision [71] and natural language processing tasks [66], e.g. image classification and action recognition [67], [70], [72], [73], image super-resolution [68], object detection [69].

Some recent works introduce attention mechanism into GAN for image synthesis [16], object transfiguration [74] or

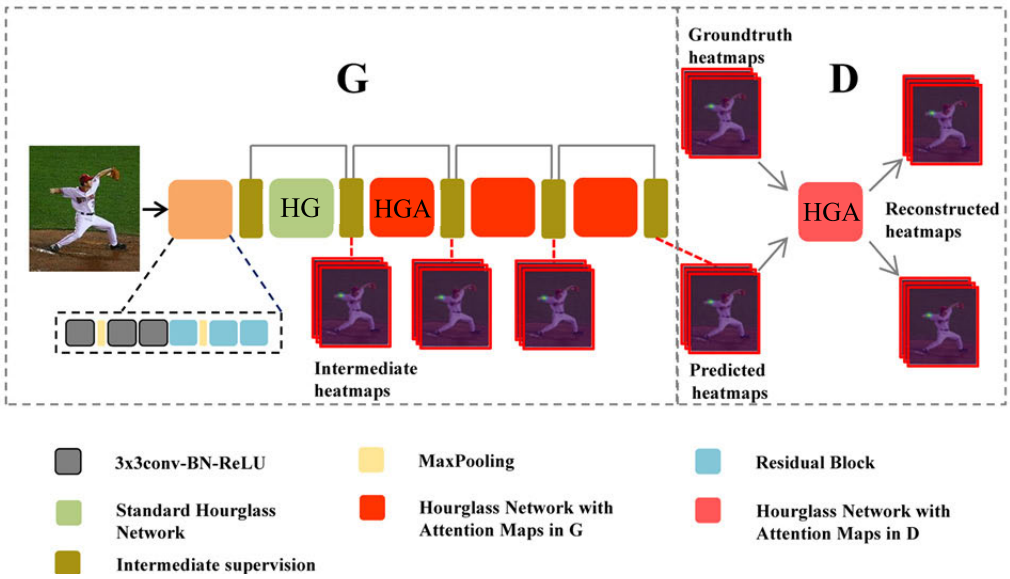


FIGURE 1. The framework of our network. HG: a single stack hourglass network, HGA: hourglass networks with self-attention, G: generator, D: discriminator.

face attribute editing [75]. In [74] and [75], attention networks lead the generator to pay important attention to specific image regions.

Zhang *et al.* [16] propose the Self-Attention GAN (SAGAN) which can model long-range dependencies for image generation tasks. The self-attention module calculates response at a position as a weighted sum of the features at all positions. Armed with self-attention, the discriminator can ensure detailed features in distant portions of the images to be consistent with each other. That is, the discriminator can more accurately enforce complicated geometric constraints on the global image structure. Note that, self-attention mechanism can learn long-range dependencies, but convolutions can only model local dependencies with local receptive fields.

III. HUMAN POSE ESTIMATION WITH SELF-ATTENTION GENERATIVE ADVERSARIAL NETWORKS

As mentioned above, although human pose estimation has been significantly advanced by deep learning, still, all the difficulties lie in occlusion, overlapping with other people, or clutter background. In such cases, the model may find similar features which belong to the background or another person. So, body structural constraints are needed. Recent works try to improve the performance of human pose estimation by refinements [10], [12], [14], which are shown to be efficient, since such refinement processes are indeed to learn structural constraints of human body joints.

The generated poses can be refined by GAN [10], [11], in which the discriminator checks the structural constraints of human body parts and distinguish implausible poses to guide and refine the generator training. But there is no attention mechanism in discriminator or generator [10]. As stated in [16], the self-attention mechanism is powerful to model long-range dependencies in the feature maps. It is complementary to convolutions, which only models local dependencies with local receptive fields. So Self-GAN [10] cannot fully learn the whole body structural constraints, which will be important in cases of occlusion, invisible joints or crowd background to ensure plausible poses.

In this paper, we introduce self-attention mechanism into Self-GAN for human pose estimation. By taking advantage of self-attention mechanism to learn long-range dependencies of body parts, the performance can be further improved.

A. THE NETWORK ARCHITECTURE

The framework is illustrated in Fig. 1. The model consists of two networks, the generator G and the discriminator D. Both use Hourglass networks [8] as their backbone. Hourglass networks are fully convolutional networks constructed with residual blocks [27] and conv-deconv architecture [26]. The generator generates heatmaps that indicate the confidence score at every location for all the body joint keypoints. The discriminator reconstructs both the predicted heatmaps and the ground truth heatmaps and distinguishes real from fake ones by adversarial training.

Note that, in Fig. 1, both the generator G and the discriminator D contain self-attention architectures, which are indicated as red blocks. They will be illustrated in detail in the next section.

1) GENERATOR

We use Hourglass networks as the generator. Following previous works [8], [9], the input images are first warped to the same resolution of 256×256 and then fed into Hourglass network. The network starts with a 7×7 convolution layer with stride 2, followed by a residual module and a round

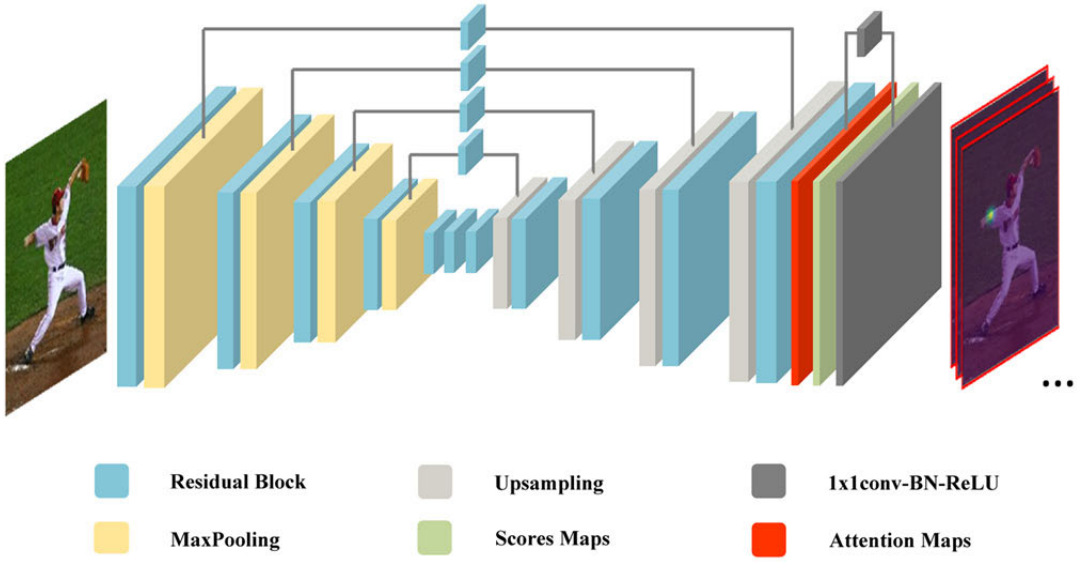


FIGURE 2. Overview of our generator framework. We show one hourglass network with self-attention architecture. The attention maps are generated by self-attention residual module (SARM) (see FIGURE 4 for more details).

of max pooling to reduce the resolution from 256 to 64. So the highest and the final output resolution is 64×64. Then, multiple hourglass modules are stacked to predict the body joint heatmaps. All residual modules output 256 features. The repeated bottom-up (from high resolutions to low resolutions), top-down (from low resolutions to high resolutions) structure together with skip connections allow processing features across all scales and capturing various spatial relationships associated with different body joints. Intermediate supervision at the end of each stack is also critical to the network's final performance.

We use 4-stack hourglass networks as generator in our experiments. The first one is the standard hourglass network, and the next there hourglass networks are integrated with attention modules. One of the hourglass networks with self-attention architectures is shown in Fig. 2. We design a new Self-Attention Residual Module (SARM) by adding the self-attention structure into residual module, as shown in Fig. 4 (a) and (b). In generator, we use the attention structure of Fig. 4 (b) that is SARM-B. Indeed we have tried several different forms of attention modules and put them at the different parts of hourglass network. Finally we adopt the attention strategy by putting the attention module at the end of the hourglass networks (see Fig. 1 and Fig. 2), which can efficiently improve the performance.

2) DISCRIMINATOR

The framework of the proposed discriminator is illustrated in Fig. 3. We use 1-stack Hourglass network as the discriminator. In standard Hourglass networks [8], the building blocks are residual modules. In this work, we introduce self-attention mechanism into the discriminator. We utilize SARMs (that is SARM-A or SARM-B, See FIGURE 4 for more details) as the skip connections to connect blocks with the same semantic meanings, that is, the blocks in bottom-up and top-down processing with the same resolution scale of feature maps.

3) VARIANTS OF SARM STRUCTURE

Formally, the SARM can be formulated as follows. Let $x \in \mathbb{R}^{C\times N\times N}$ be the input image features of the l-th layer, where C is the channel number and N the resolution of

feature map. We first transform x into feature spaces ***f, g, h*** to calculate the attention and self-attention feature maps. Let $f(x) = W_f x$, $g(x) = W_g x$ and $h(x) = W_h x$, where $W_f \in \mathbb{R}^{C'\times C}$, $W_g \in \mathbb{R}^{C'\times C}$ are the learned weight matrices which are implemented as 1×1 convolutions. As to $W_h \in \mathbb{R}^{C\times C}$, in Fig. 4(a) it is the residual module, while in Fig. 4(b) it is implemented as 1×1 convolutions. We set $C' = C/8$ in experiments.

Then the softmax operation on each row is:

$$\beta_{ji} = \frac{e^{s_{ij}}}{\sum_{i=1}^{N} e^{s_{ij}}}, \text{ where } s_{ij} = f(x_i)^T g(x_j) \tag{1}$$

β_{ji} indicates the extent to which the model attends to the ith location when generating the jth region.

In Fig. 4, we design two self-attention modules. In SARM-A (Fig. 4(a)), the output of the self-attention layer is $a = (a_1, a_2, \ldots, a_j, \ldots, a_N) \in \mathbb{R}^{C\times N\times N}$, where,

$$a_j = \sum_{i=1}^{N} \beta_{ji} h(x_i), \text{ where } h(x_i) = W_h x_i \tag{2}$$

In SARM-B (Fig. 4(b)), we multiply the output of the self-attention layer by a scale parameter γ and add to the output of residual module $f_0(x; w_{f_0})$. Then the final output of SARM-B is:

$$y = \gamma a + f_0(x; w_{f_0}) \tag{3}$$

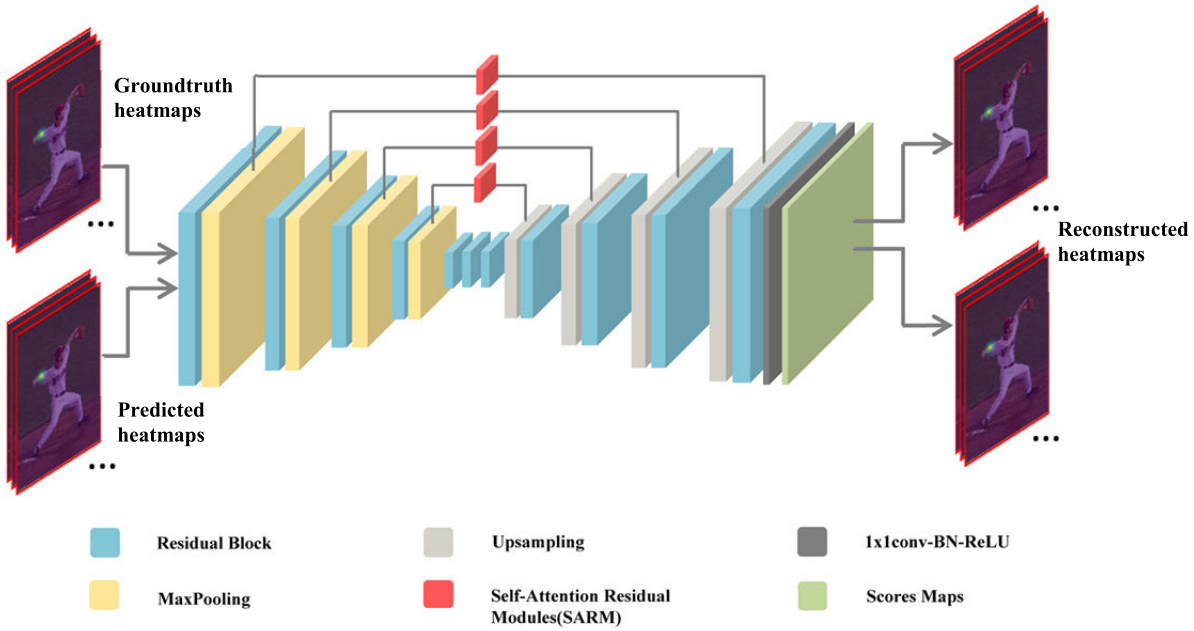


FIGURE 3. Overview of our discriminator framework. The discriminator is a single hourglass network, in which the skip connections along blocks with the same semantic meanings adopt self-attention residual modules (SARMs).

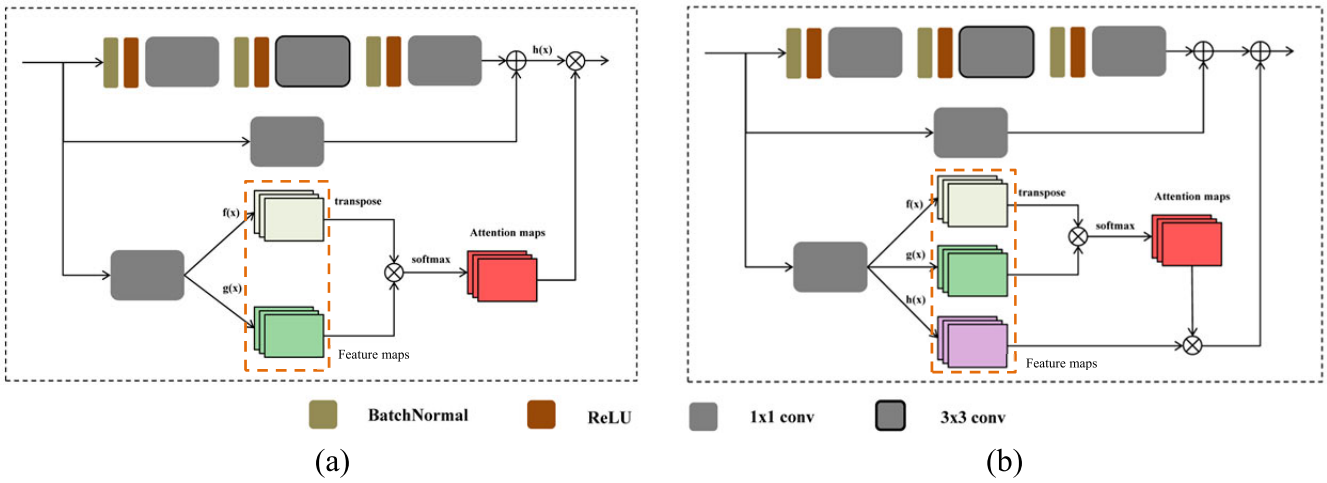


FIGURE 4. Our proposed self-attention residual modules (SARMs). (a) SARM-A produces the attention map and then matrix multiply it with the output of the residual module, while (b) SARM-B first matrix multiply the attention map with the input and then add such result to the output of the residual module. The ⊗ denotes matrix multiplication and ⊕ denotes matrix summation. The softmax operation is performed on each row.

where $\gamma \in [0, 1]$ is initialized as 0 and increased during training. The networks first rely on local cues and then gradually learn to pay more attention to long-range relationships.

Note that, in CPM [6] and Hourglass networks, long-range spatial relationships associated with body joints can be learned along layers of convolutions [6], [8], [10]. However, there are some drawbacks relying on convolutions to model dependencies across different image regions. Since convolution processes information in local receptive field, long range dependencies can only be processed through several convolutional layers. But with the increasing of layers, it will be hard to optimize parameters to coordinate multiple layers to capture dependencies. Furthermore, these parameterizations may be brittle and failing for test inputs. Thus using convolutional layers alone is computationally inefficient for modeling long-range dependencies in images.

Self-attention mechanism [16] calculates response at a position as a weighted sum of features at other locations, which is a good balance between modeling long-range dependencies and computational efficiency. So, Self-Attention can be complementary to convolutions and more suitable to capture widely separated spatial long-range multi-level

dependencies among body joints in human pose estimation problem.

The ground-truth heatmaps and generated heatmaps together with the corresponding input image are all fed into discriminator. The discriminator reproduces a new set of heatmaps. By adversarial training, the discriminator checks that features in different positions of the heatmaps are consistent with each other. Armed with self-attention, the discriminator can more accurately enforce global geometric structural constraints on generated human pose.

B. TRAINING THE SELF-ATTENTION GAN

1) TRAINING THE GENERATOR

Assume the generator consists of S stacks of hourglass modules. Each hourglass network predicts K heatmaps of size 64×64 for K body joints. We use heatmaps to represent the body joint locations. Denote the ground-truth locations as $z = \{z_k\}_{k=1}^{K}$, where $z_k = (x_k, y_k)$ indicates the location of the kth body joint in the image. Then the groud-truth heatmap is generated from a Gaussian with mean z_k and variance Σ as:

$$H_k(p) \sim \mathcal{N}(z_k, \Sigma) \tag{4}$$

where p $\in \mathbb{R}^2$ denotes the location, and Σ is set as identity matrix I.

The intermediate supervision is attached at the end of each hourglass. The loss is defined by the mean squared error:

$$L_{mse} = \frac{1}{2}\sum_{i=1}^{S}\sum_{k=1}^{K}\left\|H_{ik} - \hat{H}_{ik}\right\|^2 \tag{5}$$

where H_{ik} is the ground-truth heatmap of kth joints at the ith stack, and $\hat{H}_{ik}$ is the generated heatmap.

An adversarial loss from the discriminator is also tied to the generator:

$$L_{adv} = \frac{1}{2}\sum_{k=1}^{K}\left\|\hat{H}_k - D(\hat{H}_k, X)\right\|^2 \tag{6}$$

where $\hat{H}_k$ is the output heatmaps of the generator's last hourglass stack, D is the discriminator, and X is an input image.

The total loss for generator is:

$$L_G = L_{mse} + \lambda_G L_{adv} \tag{7}$$

where $\lambda_G \in [0, 1]$ is a hyperparameter to balance the MSE and adversarial loss. In our experiment, λ_G was set to 0.1.

Training the generator is done by back-propagating the loss L_{mse} from generator itself and the adversarial loss L_{adv} from the discriminator. The adversarial loss enables the generator to produce plausible human body poses.

2) TRAINING THE DISCRIMINATOR

The discriminator is in charge of distinguishing real poses from generated ones. Inputs to the discriminator include both heatmaps predicted by the generator and the ground truth heatmaps. The discriminator will reconstruct new heatmaps. Losses for generated and ground truth heatmaps are defined as:

$$L_{gen} = \frac{1}{2}\sum_{k=1}^{K}\left\|\hat{H}_k - D(\hat{H}_k, X)\right\|^2 \tag{8}$$

$$L_{real} = \frac{1}{2}\sum_{k=1}^{K}\left\|H_k - D(H_k, X)\right\|^2 \tag{9}$$

Algorithm 1 The Training Procedure

Generator Input: An image X
Discriminator Input: Generated heatmaps H and the corresponding ground-truth heatmaps $\hat{H}$
Repeat
(1) Forward discriminator by $D(H, X)$
(2) Compute gradient ∇f_D by Eq. (9)
(3) Forward generator by $\hat{H} = G(X)$
(4) Compute gradient ∇f_G by Eq. (5)
(5) Forward discriminator by $D(\hat{H}, X)$
(6) Accumulate gradient ∇f_D by Eq. (8)
(7) Backpropagation of ∇f_D through discriminator
(8) Accumulate gradient ∇f_G by Eq. (6)
(9) Backpropagation of ∇f_G through generator
Until $\hat{H}$ stops improving.

During training, the discriminator will minimize L_{real} while maximize L_{gen} in an adversarial style. So the total discriminator loss is:

$$L_D = L_{real} - \lambda_D L_{gen} \tag{10}$$

$\lambda_D \in [0, 1]$ is an equilibrium term to balance the discriminator and generator, which can be dynamically tuned during training [49]. In our experiment, λ_D was set to 0.5.

C. INFERENCE

At inference time, testing images should also be warped to the same resolution of 256×256 as that for training images. We use the well trained generator to predict the final human poses. We obtain the predicted body joint locations $\hat{z}_k$ from the predicted heatmaps generated from the last stack of hourglass by taking the locations with the maximum score:

$$\hat{z}_k = \underset{p}{argmax}\, \hat{H}_k(p), \quad k = 1, \ldots, K \tag{11}$$

The final prediction of the network is the max activating location of the heatmap for a given joint. The predicted location is relative to 64×64 resolution of output heatmaps. Finally, we transform the locations back to the original coordinates in the input images.

IV. EXPERIMENTS

A. DATASETS AND IMPLEMENTATION DETAILS

We evaluate our method on two widely used human pose estimation benchmarks, Leeds Sports Pose (LSP) [76] and MPII Human Pose Dataset [22].

TABLE 1. Comparisons of PCK@0.2 scores on the LSP test set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Lifshitz et al. [78], ECCV'16	96.8	89.0	82.7	79.1	90.0	86.0	82.5	86.7
Pishchulin et al. [32] , CVPR'16	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov et al. [33], ECCV'16	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei et al. [6], CVPR'16	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat et al. [79], ECCV'16	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al. [25], CVPR'17	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Yang et al.[9], ICCV'17	98.3	94.5	92.2	88.9	94.7	95.0	93.7	93.9
Ning et al.[80], TMM'17	98.2	94.4	91.8	89.3	94.7	95.0	93.5	93.9
Chou et al.[10], arXiv'17	**98.2**	**94.9**	**92.2**	**89.5**	**94.2**	**95.0**	**94.1**	**94.0**
Ours-A	**98.6**	**95.1**	92.1	**89.8**	**94.7**	94.9	93.9	**94.2**
Ours-B	**98.5**	**94.9**	**92.5**	**89.8**	**94.6**	**95.2**	93.9	**94.2**
Ours-C	98.3	**95.1**	**92.6**	**89.8**	**94.6**	**95.0**	**94.5**	**94.3**

The LSP and its extended dataset [76] contain total 12k images with poses in various sports. 11000 images are used for training and 1000 for testing. Each image is annotated with 14 keypoint locations. The center and scale of annotated person are calculated to be used in training. The dataset is challenging because of its noisy labels and various poses.

MPII dataset [22] contains about 25k images and over 40k annotated people, which covers a wide range of human activities. Each image may contain multiple persons. We focus on single person pose estimation. There exist some missing annotations for some persons. We just consider persons with full annotations. We follow previous works [9] to split the training set into train and validation subsets. Train set has 14679 images with 22246 persons, validation set has 2729 images with 2958 persons, and test set has 6619 images with 11731 persons. The test set has no annotations. Each person is annotated with 16 joints, the center and scale.

We do data augmentation following previous works [8]–[10]. All input images are 256×256 cropped from resized images according to the annotated human body centers and scales. An input image is randomly flipped horizontally, rotated by an angle in [-30, 30] degrees, and scaled with factors in [0.75, 1.25].

We implement our models using Torch7 [77] deep learning libraries. The networks are optimized by RMSprop algorithm with a batch size of 6 for 200 epochs. Training is performed on a server with 16GB NVIDIA Tesla P100 GPU. The learning rate is initialized as 1×10^{-4} and dropped by 5 at 40*th*, 60*th* and 80*th* epoch. Our model takes about 4 days to train on the training set.

B. EVALUATION METRICS

We use Percentage of Correct Keypoints (PCK) [18] to measure performance on LSP dataset, and use PCKh [22] on the MPII dataset.

Let $\hat{z}_k$ be the predicted location of the kth body joint, z_k is the corresponding ground truth location, then PCK is defined as:

$$\frac{\left\| z_k - \hat{z}_k \right\|_2}{\left\| z_{lhip} - z_{rsho} \right\|_2} \leq r \tag{12}$$

where z_{lhip} and z_{rsho} denote the ground truth locations of the left hip and right shoulder, respectively. $r \in [0, 1]$ is a threshold. PCK calculates the percentage of correct detections that fall within a normalized distance. Here the distance is normalized by the torso size.

For MPII, the distance is normalized by a fraction of head size, and the matching threshold is defined as 50% of the head segment length. Such metric is referred to as PCKh. PCKh uses head size to make the metric articulation independent [22].

C. RESULTS

1) QUANTITATIVE RESULTS

In our experiments, we implement two self-attention strategies, SARM-A and SARM-B, as shown in Fig. 4 (a) and (b), and we apply them both in generator and discriminator. Firstly, we just integrate the self-attention modules into the discriminator (Fig. 3) while leaving the generator be standard hourglass networks. In such cases, we denote our method with the attention structure of SARM-A (Fig. 4 (a)) in discriminator as Ours-A, and the attention structure of SARM-B (Fig. 4 (b)) in discriminator as Ours-B. Then we further integrate the self-attention modules into the generator. We denote our method with SARM-B both in the generator (Fig. 2) and discriminator (Fig. 3) as Ours-C.

Comparisons of the PCK scores at the threshold of 0.2 (PCK@0.2) on the LSP test set is presented in Table 1. Following previous works [9], [10], our models are trained when adding MPII training set to the LSP training and LSP extended training set. Both of our methods, Ours-A and Ours-B, achieve the performance of 94.2%, and improve the previous best result [10] by 0.2%. Ours-C achieves the best performance of 94.3%, and improves the previous best result [10] by 0.3%. Ours-A achieves the best scores on four body parts, head, shoulder, wrist and hip, with 98.6%, 95.1%, 89.8% and 94.7% respectively. Ours-B achieves better scores on two body parts, wrist and knee, with 89.8% and 95.2% respectively. Ours-C achieves the new best scores on four body parts, shoulder, elbow, wrist and ankle, with 95.1%, 92.6%, 89.8% and 94.5% respectively.

TABLE 2. Comparisons of PCKh@0.5 scores on the MPII test set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Pishchulin et al. [32], CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. [78], ECCV'16	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary et al. [81], ECCV'16	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al. [82], BMVC'16	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagiannis et al. [83], FG'17	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutdinov et al. [33], ECCV'16	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [6], CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat et al. [79], ECCV'16	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. [8], ECCV'16	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Tang et al. [30], ECCV'18	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning et al. [80], TMM'17	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu et al. [25], CVPR'17	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al.[10], arXiv'17	**98.2**	**96.8**	**92.2**	**88.0**	**91.3**	**89.1**	**84.9**	**91.8**
Chen et al.[11], ICCV'17	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang et al.[9], ICCV'17	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al.[31], ECCV'18	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Ours-A	**98.5**	**96.7**	92.5	**88.7**	91.0	88.8	85.9	92.0
Ours-B	**98.5**	**96.7**	92.3	**88.7**	91.1	89.0	86.0	**92.1**
Ours-C	98.0	**96.8**	**92.6**	**88.8**	**91.4**	**89.4**	**86.7**	**92.3**

* The red numbers indicate the best scores, and the blue ones are second.

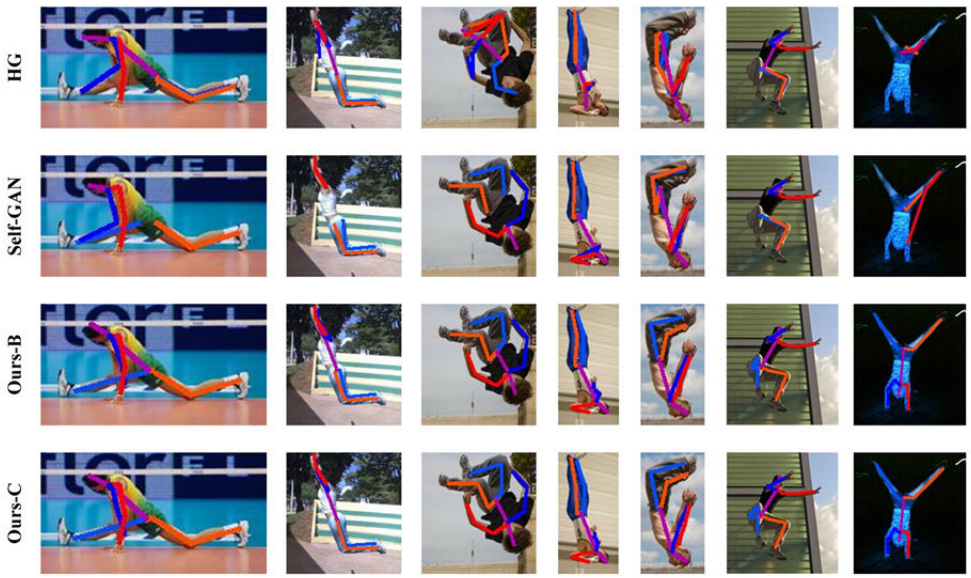


FIGURE 5. Qualitative comparisons on LSP.

In Table 2, we present the comparisons of the PCKh scores at the threshold of 0.5(PCKh@0.5) on the MPII test set. Our approaches Ours-A achieves 92.0%, Ours-B improve the performance to 92.1%, and Ours-C improve the performance to 92.3% which are comparable to previous state-of-the-art results [9], [31]. In addition, Ours-A and Ours-B both achieve the best scores on the body part of head, and Ours-C achieves the new best scores on four body parts, shoulder, wrist, hip and ankle, with 96.8%, 88.8%, 91.4% and 86.7% respectively. Compared with the counterpart method [10] whose total PCKh@0.5 is 91.8%, our method acquires 0.5% improvement by taking advantage of self-attention mechanism. Specifically, our methods achieve 0.3%, 0.4% and 0.3% improvements on head, elbow and knee respectively. For the most challenging parts to be detected as wrist and ankle, our improvements are even notable, with **0.8% and 1.8%** respectively.

2) QUALITATIVE COMPARISONS

We show some qualitative comparison results on LSP in Fig. 5, and on MPII in Fig. 6. We compare our models with other counterpart methods. In each figure, the first row contains some results predicted by Hourglass (HG) [8], the second row by Self-GAN [10], and our results are in the third and fourth row.

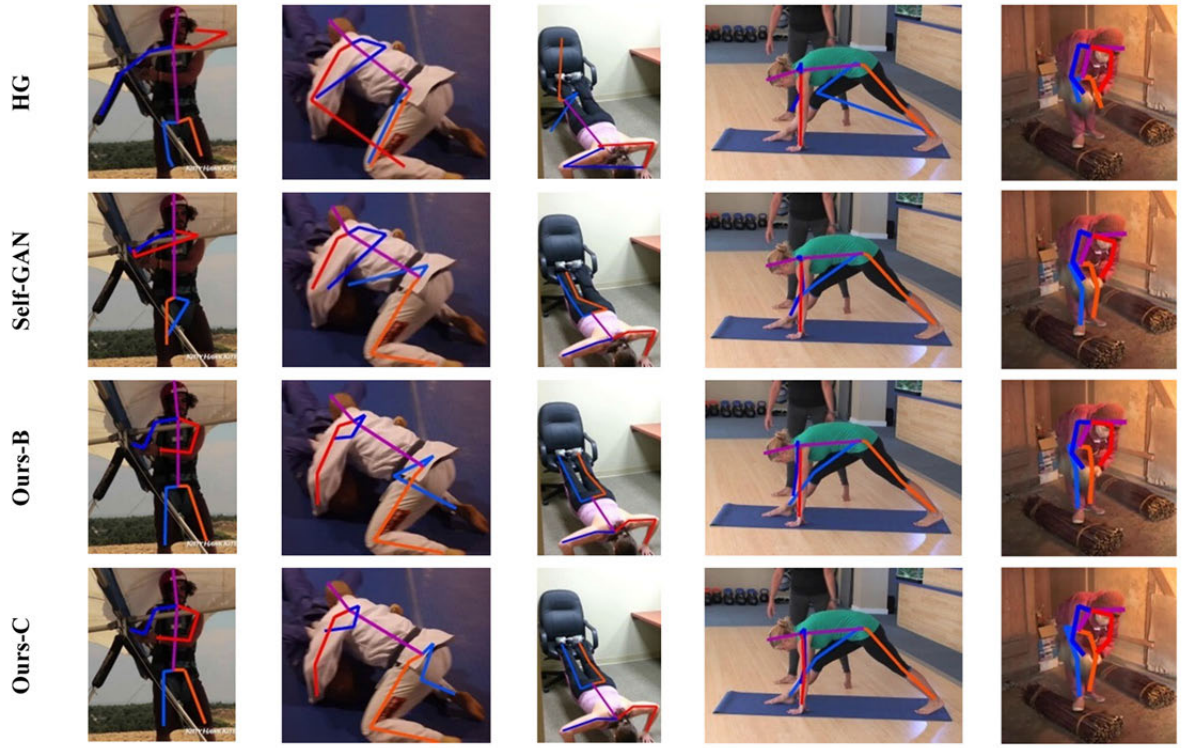


FIGURE 6. Qualitative comparisons on MPII.

TABLE 3. Ablation study on the LSP test dataset.

Experimental Settings	Description	Generator (G)	Discriminator (D)	PCK@0.2 scores (Mean)
1	Standard 4-stacked HG		×	93.6
2	SARM-B in one HG		×	93.7
3	SARM-B in two HG		×	93.9
4	SARM-B in three HG		×	**94.0**
5	Self-GAN [10]			94.0
6	Ours-B			94.2
7	Ours-C			**94.3**

We can see that, by utilizing adversarial training in GAN framework, the generator can be refined to correct some errors to produce more plausible poses [10]. Further, we introduce attention mechanism into the generator and the discriminator to enhance global long-range dependencies among body joints. The performances are obviously improved.

D. ABLATION STUDY

We do some ablation studies to verify the effectiveness of our designed self-attention modules. Table 3 contains ablation study on the LSP test dataset. We explore total seven experimental settings. Experiments 1-4 are just based on 4-stacked Hourglass Networks (HG). Experiment 1 is the standard 4-stacked HG, without self-attention mechanism. The mean PCK@0.2 score is 93.6%. From experiment 2 to 4, we integrate the self-attention module, that is SARM-B (Fig. 4(b)), into one HG, two HG, and three HG, step by step. We can see that the mean PCK@0.2 scores are also progressively improved from 93.7% to 93.9% and then 94.0%.

Experiments 5-7 are based on the framework of GAN, where the generator is 4-stacked HG and the discriminator is one HG. Experiment 5 is the Self-GAN [10], without self-attention mechanism. Comparing experiment 5 with experiment 1, we can see that by utilizing GAN, the score is

TABLE 4. Ablation study on the MPII validation dataset.

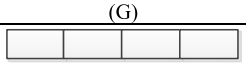
Experimental Settings	Description	Generator (G)	Discriminator (D)	PCKh@0.5 scores (Mean)
1	Standard 4-stacked HG		×	87.22
2	SARM-B in one HG		×	87.35
3	SARM-B in two HG		×	87.40
4	SARM-B in three HG		×	87.43
5	Ours-C			**88.07**

TABLE 5. Ablation study on the MPII test dataset.

Experimental Settings	Description	Generator (G)	Discriminator (D)	PCKh@0.5 scores (Mean)
1	Self-GAN [10]			91.8
2	Ours-B			**92.1**
3	Ours-C			**92.3**

*  : One standard HG (Hourglass Networks)  : One HG with Self-Attention Modules
SARM: Self-Attention Residual Modules ×: No Discriminator

TABLE 6. Comparisons of parameter, moder size and average runtime.

Method	Parameters	Model size	Runtime(s) on MPII test set	Runtime(s) on LSP test set
Newell et. al. [8]	23.7M	196.08M	1.204	0.227
Chou et al.[10]	24.7M	191.6M	1.149	0.202
Ours-C	25.46M	196.21M	1.266	0.239

improved from 93.6% to 94.0%, which indicates the effectiveness of GAN. And experiment 4 has the same score of 94.0% as experiment 5. Experiment 6 and 7 are exactly our methods of Ours-B and Ours-C (see Section 4.3 for detail), which are implemented by gradually applying self-attention module SARM-B into discriminator and generator, and the scores are also increased little by little from 94.2% to 94.3%. It is clear that the self-attention mechanism can further improve the performance of GAN.

Table 4 and Table 5 are our ablation studies on MPII validation and test dataset respectively. All our results verify that: (1) the self-attention mechanism can improve the performance of hourglass networks for human pose estimation; (2) the performance can be further improved by integrating the self-attention modules into the framework of GAN.

Table 6 compares the complexity and runtime of Our-C model with other counterpart approaches.

V. CONCLUSION

In this paper, we propose to utilize self-attention GAN to further improve the performance of human pose estimation. In the GAN framework, the generator and discriminator are all stacked Hourglass networks. The generator is responsible for predicting poses, while the discriminator acts to enforce structural constraints of human body joints to refine the poses. Since modeling long-range dependencies among image regions just by convolutions with local receptive fields is not efficient and enough, so we further introduce self-attention mechanism into the generator and the discriminator.

The self-attention mechanism allows modeling long-range dependencies among body joints. With self-attention, the generator can pay more attention to salient body joints, while the discriminator can check that body joints in distant portions of the body are consistent with each other. So entire body joints geometry constraints can be further enforced during training of the generator, which will be important in cases of occlusion, invisible joints or crowd background to ensure plausible poses.

Experiments on two standard human pose estimation benchmarks demonstrate the effectiveness of our approach.

REFERENCES

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 7444–7452.

[2] Z. Huang, C. Wan, T. Probst, and L. V. Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Jul. 2017, pp. 6099–6108.

[3] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Jun. 2018, pp. 5323–5332.

[4] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Jun. 2014, pp. 1653–1660.

[5] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. NIPS*, 2014, pp. 1799–1807.

[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Jun. 2016, pp. 4724–4732.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Jul. 2017, pp. 7291–7299.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, Oct. 2016, pp. 483–499.

[9] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1281–1290.

[10] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 17–30.

[11] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1212–1221.

[12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. CVPR*, Jun. 2018, pp. 7103–7112.

[13] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. ECCV*, Sep. 2018, pp. 466–481.

[14] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. ECCV*, Sep. 2018, pp. 190–206.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[16] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, May 2019, pp. 1–10.

[17] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 3487–3494.

[18] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[19] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. CVPR*, Jun. 2008, p. 7.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[22] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Jun. 2014, pp. 3686–3693.

[23] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 740–755. [Online]. Available: https://github.com/pdollar/coco

[24] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 3073–3082.

[25] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, Jul. 2017, pp. 1831–1840.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[28] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, J. Malik, B. E. Shi, and S. Savarese, "Feedback networks," in *Proc. CVPR*, Jul. 2017, pp. 1308–1317.

[29] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.

[30] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-Nets for efficient landmark localization," in *Proc. ECCV*, Sep. 2018, pp. 339–354.

[31] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 713–728.

[32] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 4929–4937.

[33] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, Oct. 2016, pp. 34–50.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Oct. 2017, pp. 2961–2969.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Sep. 2014, pp. 1–14.

[36] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 2334–2343.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, C.-Y. Fu, A. C. Berg, and S. Reed, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 21–37.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[39] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proc. CVPR*, Jun. 2018, pp. 5167–5176.

[40] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proc. CVPR*, Jun. 2018, pp. 350–359.

[41] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, and L. Lin, "LSTM pose machines," in *Proc. CVPR*, Jun. 2018, pp. 5207–5215.

[42] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H. P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," in *Proc. ACM Trans. Graph. (SIGGRAPH)*, Jul. 2017, p. 44.

[43] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D human pose estimation in the wild by adversarial learning," in *Proc. CVPR*, Jun. 2018, pp. 5255–5264.

[44] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.

[45] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, Nov. 2016, pp. 1–16.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, Jul. 2017, pp. 214–223.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[48] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. ICLR*, Sep. 2017, pp. 1–17.

[49] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*. [Online]. Available: https://arxiv.org/abs/1703.10717

[50] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. ICLR*, Feb. 2018, pp. 1–26.

[51] H. Liu, X. Gu, and D. Samaras, "A two-step computation of the exact GAN Wasserstein distance," in *Proc. ICML*, Jul. 2018, pp. 3159–3168.

[52] N. Lei, K. Su, L. Cui, S.-T. Yau, and D. X. Gu, "A geometric view of optimal transportation and generative model," 2017, *arXiv:1710.05488*. [Online]. Available: https://arxiv.org/abs/1710.05488

[53] N. Lei, Z. Luo, S.-T. Yau, and D. X. Gu, "Geometric understanding of deep learning," 2018, *arXiv:1805.10451*. [Online]. Available: https://arxiv.org/abs/1805.10451

[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2223–2232.
[55] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in *Proc. ECCV*, Sep. 2018, pp. 119–135.
[56] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR*, Jun. 2018, pp. 8789–8797.
[57] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. ECCV*, Sep. 2018, pp. 818–833.
[58] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Proc. NIPS*, Aug. 2018, pp. 1–14.
[59] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, Jul. 2017, pp. 4681–4690.
[60] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," 2018, *arXiv:1806.03589*. [Online]. Available: https://arxiv.org/abs/1806.03589
[61] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. ECCV*, Sep. 2018, pp. 206–221.
[62] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," in *Proc. NIPS*, 2018, pp. 474–484.
[63] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. NIPS*, 2018, pp. 1222–1233.
[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
[65] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *Proc. ICLR*, Mar. 2017, pp. 1–15.
[66] M. Barrett, J. Bingel, N. Hollenstein, M. Rei, and A. Søgaard, "Sequence classification with human attention," in *Proc. CoNLL*, Oct. 2018, pp. 302–312.
[67] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," in *Proc. NIPS*, 2018, pp. 352–361.
[68] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, Sep. 2018, pp. 286–301.
[69] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, pp. 3–19.
[70] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. ECCV*, Sep. 2018, pp. 373–389.
[71] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS*, 2014, pp. 2204–2212.
[72] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "stagNet: An attentive semantic RNN for group activity recognition," in *Proc. ECCV*, Sep. 2018, pp. 101–117.
[73] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *Proc. ECCV*, Sep. 2018, pp. 135–151.
[74] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-GAN for object transfiguration in wild images," in *Proc. ECCV*, Sep. 2018, pp. 164–180.
[75] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proc. ECCV*, Sep. 2018, pp. 417–432.
[76] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, Aug. 2010, p. 5.
[77] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like environment for machine learning," in *Proc. NIPS Workshop*, 2011, pp. 1–6.
[78] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 246–260.
[79] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV*, Oct. 2016, pp. 717–732.
[80] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.
[81] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. ECCV*, Oct. 2016, pp. 728–743.
[82] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe, "An efficient convolutional network for human pose estimation," in *Proc. BMVC*, Sep. 2016, p. 2.
[83] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, May/Jun. 2017, pp. 468–475.
[84] R. Wang, Z. Cao, X. Wang, Z. Liu, and X. Zhu, "Human pose estimation with deeply learned multi-scale compositional models," *IEEE Access*, vol. 7, pp. 71158–71166, 2019.